Visualizing and Understanding Ranking Trends of Wikipedia Top Viewed Pages

Roy G. Biv*
Starbucks Research

Ed Grimley†
Grimley Widgets, Inc.

Martha Stewart‡
Martha Stewart Enterprises
Microsoft Research

Figure 1: By exploring the WikiRankVis system, we found the pattern of weekly TV series. The rank of page "Glee (TV series)" (an American TV series) goes up every Wednesday, indicating the day when Glee is on show, and goes down after. By exploring the page's similar pages, we get other pages of Glee, and other weekly TV series such as "The Big Bang Theory" and "Two and a Half Men" in the page link network. The rank line chart shows the weekly pattern clearly.

Abstract

Wikipedia top page view statistics are collections of top viewed Wikipedia pages over time, and of great importance in analyzing viewers' interest in current affairs. However visualizing Wikipedia page ranking trends usually suffers from great visual clutter, which is also a common challenge in large time series visualization. Following the gestalt's law of continuity, we tried out a variety of visual designs of top ranked pages and evaluated their efficiency in describing ranking trends. Based on the visual designs we implemented **WikiTopReader**, a reader of Wikipedia page rank. Users are able to explore connections among those top viewed pages by connecting the page rank behavior with the page link information. Such combination enhances the unweighted Wikipedia page link network and brings users' page of interest to broader attention. The evaluation result shows the feasibility of the visual design and the system.

Keywords: Wikipedia page view, Wikipedia page link, rank time series, visualization.

*e-mail: roy.g.biv@aol.com
†e-mail:ed.grimley@aol.com
‡e-mail:martha.stewart@marthastewart.com

1 Introduction

Wikipedia is considered as the biggest online encyclopedia, whose everyday page view throughput can be more than 600M in all languages [1]. It has become a knowledge exchange platform where users learn and contribute their knowledge. Due to the sustained large scale of knowledge accumulation, Wikipedia also becomes a huge and growing knowledge warehouse. It has accumulated over 30M pages and more than 20M "editors" are maintaining the pages.

It would take more than human lifetime to digest all the knowledge gathered in Wikipedia. But if we just want to catch up with the period of time and understand what your friends are talking about, we would just read through top viewed pages. According to the 20/80 principle, the ranking data of Wikipedia top page view statistics (**Wikipedia page rank** for short) reflects users' major interests in Wikipedia or furthermore in current affairs. Time series of Wikipedia page rank, or **Wikipedia page ranking trends**, therefore indicates how users' social interest evolving over time.

There has been enormous work of Wikipedia contents in the infrastructural research field such as NLP and RDF database. However in terms of analyzing Wikipedia contents, data mining strategies does not satisfy users' various needs towards the long Wikipedia page ranking trends. On the other hand, visualization

[1] Wikipedia article traffic. `http://stats.grok.se`

solutions usually fall into two hazards: either too simple to reveal rank trends or too chaotic to read individual page clearly. Simple solutions such as page labels requires too much user interactions to explore the whole dataset. Chaotic solutions, such as band- or river-like designs, suffer greatly from visual clutter, for which overwhelming crossings caused by such visual metaphors of continuity are most likely to blame.

To address the challenges, we designed the WikiTopReader system, which not only visualizes Wikipedia top page view statistics, but also constructs a semantic network based on a given Wikipedia page of interest. Our principle is that any good design should connect the same Wikipedia page over time without causing unnecessary perceptual complexity. Our visualization design dodges visual clutter effectively by breaking band- or river-like visualization into scattered glyphs while keeping users' perceptual continuity towards certain ranking items. In addition, we construct a semantic network that associates Wikipedia pages of similar ranking trends into a current affair. Moreover, it also characterizes a user-aware network of the original planar Wikipedia page link network. Although only practiced on the Wikipedia dataset, such design can be applied to other time series such as stock prices. In summary, we summarize our contributions as follows:

- Three glyph designs that portrays Wikipedia page ranking trends without causing visual burden,
- A mashup of the page view dataset and the page link datasets that enhances users' understanding of the page rank relations.

The rest of this paper is organized as follows. Related work is reviewed in Section 2. Section 3 describes all the tryouts of the Wikipedia page rank representation. Section 4 explains how users connect page ranking trends with the page link relations. Case studies are elaborated in Section 5, followed by a user study and its discussions in Section 6. Finally we conclude the paper in Section 7.

2 Related Work

We first review previous research work of the collections of Wikipedia datasets, then go over work on visualizing rank time series.

2.1 the Wikipedia Dataset Collections

Many researchers are trying to connect Wikipedia pages with meanings of text. Explicit Semantic Analysis (ESA) [4] is a machine learning method that assesses relatedness of textual concepts derived from Wikipedia. It is applied to answer questions like "how related are cats and mouse". The texts are represented as a weighted vector of Wikipedia-based concepts and the relatedness is regarded as a conventional metrics in low-dimensional space. Also with machine learning method, Milne and Witten's work [2] is about identifying significant terms within unstructured text. The resulting disambiguator and link detector can perform with high precision. In application, the disambiguator can enrich text phrases in other documents with links to the appropriate Wikipedia articles. And the link detector connects terms in a way users can build up the semantic structure of the document more efficiently. DBPedia [6] is a large scale multilingual knowledge base extracted from Wikipedia. Researchers extracted structured information from Wikipedia and organized them with RDF documents. Pages are connected to each other with ontology primitives such as person A is a "relative" to person B or A is a "product" of company B.

2.2 Visualization of Rank Time Series

Visualization of rank time series falls into three categories: low-dimensional embedding, curve representation and glyph representation. Low-dimensional embedding for visualizing rank time series is to embed each individual rank in a low-dimensional space. The ranks then are encoded with a heatmap [9], dots [8] and click-stream sequences[5].

Existing representation of rank items as lines, bands or rivers can often cause heavy visual clutter if there are a large number of items. To present rank trends with spirals, RankClock [7] employed radial coordinates to represent the rankings of different time points, however it can hardly scale to long rankings. The "table-graphic" [3] by Edward Tufte, or now known as slopegraph, connects time-varying items with lines to reveal temporal changes of item values. The slopegraph can also present rank time series sorting items with rank. If the rank evolution of each ranked item is considered, it can be visualized as a ranking band [2]. A similar visualization is achieved in [3], which depicted the variation of the top-50 ranked items one by one and connected the same items with curves.

Even the simplest scatter plots assisted with intuitive user selection operations [4] can be applied to explore temporal trends of ranked items. By emphasizing the subtle rank changes, RankExplorer [1] adaptively segmented the Wikipedia rank time series into groups, and extends ThemeRiver with color bars and glyphs for rank time series to convey adjacent rank changes and overall trends. It emphasizes the overall rank changes between successive time points by grouping ranked items of similar ranking orders into multiple segments at each time point. Although this method successfully depicts the rank changes within and across multiple categories, it is hardly capable of disclosing pattern of individual items, let alone complicated correlations of rank changes.

3 Various Mutants of Page Rank Representation

Our general goal of visualization is to describe the involving local ranking trend of a specific page. Based on the principle that any good description should not cause unnecessary perceptual complexity, we give up the obscure low-dimensional embedding and the overwhelming curve representation. The only alternative left is the glyph representation. Despite of its limitation, it is still required to disclose pattern of individual pages, as well as the correlations of ranking trends with the glyph representation. In the visualization of page rank time series, we followed the gestalt law of continuity and made several design proposals as follows. And section 4 describes how to reveal page correlations. Section 6 narrates a user study conducted to evaluate the best representation of page ranking trends.

3.1 The Sparkline Visualization

The most intuitive solution of time series would be sparklines of local ranking trends of pages (see Figure 2), which is widely used in stock price display or click rate display. Users would be able to see a rough local trend of any given page and make comparisons among sparklines.

However, the drawback is also obvious. It becomes so visually unpleasant when it comes to a screen full of worm-like sparklines. Also, due to limited display space, subtle changes cannot be accurately told and users do not know what is the exact next rank of the page.

3.2 The Badge Visualization

Usually lines or arrows that connect to its next position of the page can lead users attention and provide a continuous perceptual experience to users. Guided by the gestalt law of continuity, we get rid of the lines or streams that cause visual clutter but retain the page rank glyph meanwhile. The glyph we designed here is called **badge**. A

[2]Fortune-500 visualization. http://in.somniac.me/2010/01/fortune-500-visualization/

[3]WikiTop-50 visualization. http://www.chrisharrison.net/index.php/Visualizations/WikiTop50

[4]Fortune-500 visualization. http://fathom.info/fortune500/

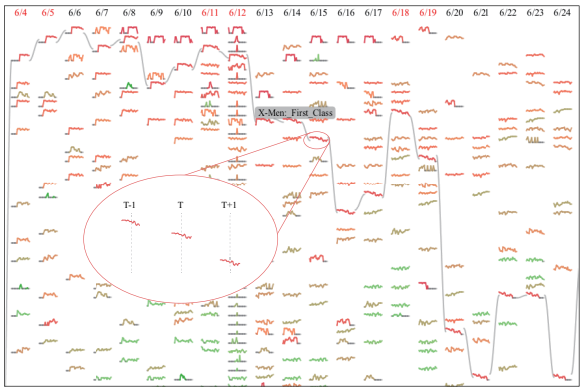

Figure 2: The sparkline representation of page ranks. Users may check any single glyph to see the page "X-Men: First Class"'s local involving trend.

badge is a simple glyph representation of page rank with two edges (see Figure 3): one pointing to its last rank position and the other pointing to its next rank position. Both visual channels, shape and color, enhance users' cognition of the same page.

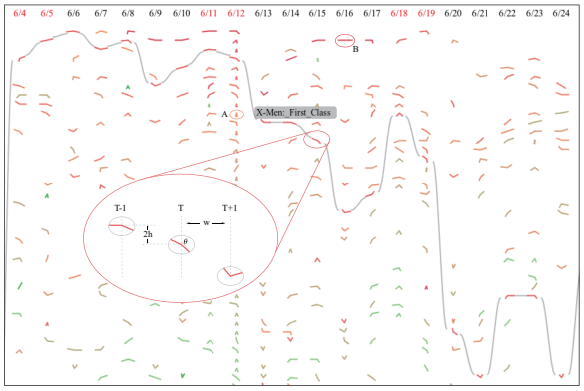

Figure 3: The badge representation of page ranks. By following the color and the directions of the two edges, users can see the involving ranking trend of "X-Men: First Class". Meanwhile, glyphs of page A and page B show two different rank involving patterns.

To maintain a consistent shape of badge, we constraint the area for a badge in an ellipse with the same width and height. The rank difference and the time interval makes an angle θ.

$$\theta = \begin{cases} \pi - \arctan(h * |d_{rank}|/w) & \text{left} \\ -\arctan(h * |d_{rank}|/w) & \text{right} \end{cases},$$

where d_{rank} is the rank difference, w and h are the width and height of the display area. Thus the badge can be represented with two line segments from the center to the edge of the ellipse respectively. The coordinates of the two point on the edge can be represented with polar coordinates as follows:

$$\begin{cases} x = w' * \cos\theta, \\ y = -h' * \sin\theta \end{cases}$$

where $2w'$ and $2h'$ are the width and height of the ellipse and are usually a little smaller than the display area.

The shape of badge, or the size of angle, naturally depicts the general ranking trend of the page. As it shows in Figure3, page A is more likely to be a new top page, flashing in and flashing out of the top page rank list. While page B is a popular page with steady involving pattern during that period. But following B's trend, it drops the day after its next day.

3.3 The Color Scheme

We designed the color scheme based on one day. Once users select a day, all pages in other days will be colored the same with that of the selected day. All pages not in the top rank list of the selected day will be filtered. It further assists users to locate a certain page with both glyph and color. To highlight those days with fewer pages left, we make a little modification of the badge visualization. Instead of coloring the glyph, we have the background colored (see Figure 4) to reveal how pages on the selected day survived in other days. Meanwhile the blank space indicates that more new pages appear on the corresponding day.

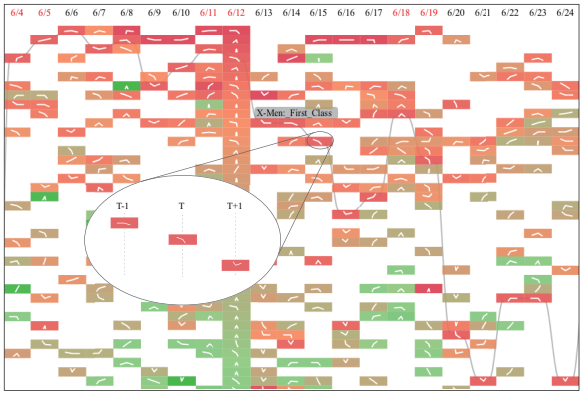

Figure 4: The badge representation of page ranks but with background colored. Users can also notice how pages on June 12th survived on the rank list.

4 Semantic Exploration

We explore the page-wise semantics with two integrated relations: **the page link relation** and **the rank time series similarity relation**. The page link relation states a semantic relation explicitly, while similarity relation is more implicit. For a current affair, users tend to query several key words that are related to the affair, resulting in concurrent rank time series. Under this assumption, we infer that those pages with similar rank series are potentially related. However, users should note that page with similar rank series are not necessarily linked. We try to enhance the planar page link network with user behaviors and capture users' limited attention to a more condense network.

4.1 Finding Pages of Similar Trend

We evaluate the dissimilarity between two pages with the proximity of their associated rank time series, under the assumption that two Wikipedia pages are potentially but not necessarily correlated if they share similar trends during their recent history. We first take all rank series as polylines and compare their similarity with a curving matching method.

We define the dissimilarity of two ranked pages (A and B) at a time point mainly based on the curve matching factor f_{cm}. The curve matching factor is the main factor for rank series similarity obtained by a curve matching method [10] which applies dynamic time warping in curve matching. The paper takes similarity of two polygonal chains equivalent to the optimal distance on the manifold made by the two chains. The optimal distance can be calculated with dynamic time warping. In addition, to get a more accurate results, it interpolates steiner points on the edges of the manifold

patches. To adapt the method we take rank time series as polylines and compare their similarity with the method as described.

We also adopt the entropy-based evaluation of clustering quality described in Chen et. al's paper [11] in finding similar pages. The similarity aware entropy score calculates the entropy of a bunch of time series. Since we've already got the pairwise similarity, we iteratively search for page with the most similar rank series and compute the entropy score until the score exceeds a certain threshold. This is obviously prior to k nearest neighbors because it gets all pages whose rank series are similar enough with that of this page. We plot the local rank trend of these pages so that users can compare how and when the pages have similar patterns.

4.2 Further Exploring Semantics

The page link network further explains the relationship between pages. For all similar pages found, we construct a pairwise network based on the page link information. Current page of interest is placed in the middle while pages linked to it are connected via edges. To further attract users' attention to the current focus, pages not linked to any other pages are scattered along the boundary of the display area.

Figure 5 shows pages of similar rank trends with that of page "X-Men First Class". Among them, there are related pages such as "X-Men" and "2011 film". And the "2011 film" page connects to "Jay Baruchel" (the script writer and the main actor of film Goon) and "Transformers: Dark of the Moon".

Figure 5: The page link network of page "X-Men First Class". Pages with similar ranking trends as well as page links include .....(picture to be replaced).

4.3 Implementation

The page view statistics for Wikimedia projects maintains raw page access records for all Wikipedia projects in all languages [5]. We have collected 14-month English page view statistics dataset (from Jun. 1st to Oct. 27th in 2011 and from January to September in 2014) and generated daily top-1000 page views in a MySql data archive. To focus on current affairs, we removed the index page, the portal pages, the error pages and other specific pages from our statistical results. Visualization only shows top-50 page views while page with similar rank series are fetched among the top-1000 pages.

We also collected page links dataset via Wikipedia page link APIs [6] and maintained data persistence via Neo4j [7], a graph database.

5 Case Studies

5.1 Wikipedia top-50 query series

The first case demonstrate the usefulness of WikiTopReader on spotting unusual events, especially on spotting and accurately recognizing distinct concurrent events which share similar ranking trends.

As shown in Figure 6, on July 23rd, 2011, four surging items: Amy, 27 Club, Anders and 2011 Norway Attack suddenly showed up to the top of the rank simultaneously. They stayed top for several days, then went down rapidly and fell of the top 50 rank list. Since mostly, such an unusual changing pattern shared by 4 items may indicate unusual events, we clicked one of the items: 27 Club for further exploration.

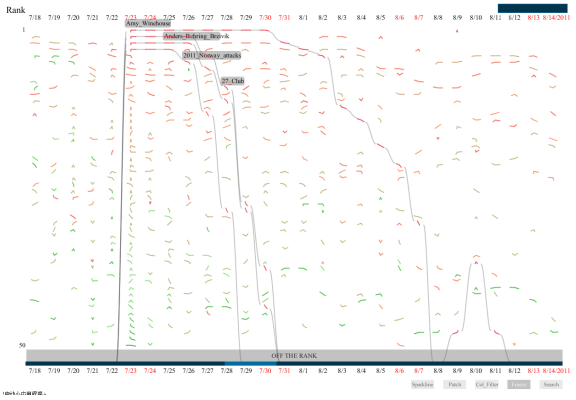

Figure 6: .

Figure 7 shows the results after 27 Club is clicked. (a) is the rank curve view of items sharing similar page view changing pattern with 27 Club, while (b) shows the network constructed by the similar items based on similarity and page-link information. (The size of each node demonstrates the similarity shared with the central node. The arc between two nodes depicts the existence of page-link relationship between the specific items represented by the two nodes.)

We could find it in (a) that most of the similar items computed by WikiTopReader shared a similar changing pattern. They suddenly surged on a specific day, stayed for several days and then went down. Hence, our system is wise in similarity computation and the curve view does make sense in showing the changing pattern.

Then, we refer to page-link network view (b) for more information. Two independent network N1, N2 are shown in the view. Amy and 27 Club are in N1, while 2011 Norway Attack and Anders are in N2. N1 and N2 are independent from each other, which means that the 4 similar items may belong to distinct events.

For detailed exploration, we filtered out the nodes without page-link information and constructed the network only with the rest. Then we got the result shown in Figure 8.

N1 and N2 are obviously independent events, but they got high attention simultaneously. N1 is about Amy and 27 Club, and N2 is about Norway and Anders. N1 could be further partitioned into N11 and N12, which are independent from each other but share the similar relation with 27 Club. Then, with the help of the Internet, we got to know that 27 Club is a team that refers to a number of

[5]Page view statistics for Wikimedia projects. http://dumps.wikimedia.org/other/pagecounts-raw/

[6]MediaWiki API. http://en.wikipedia.org/w/api.php

[7]Neo4J, a graph database. http://neo4j.com

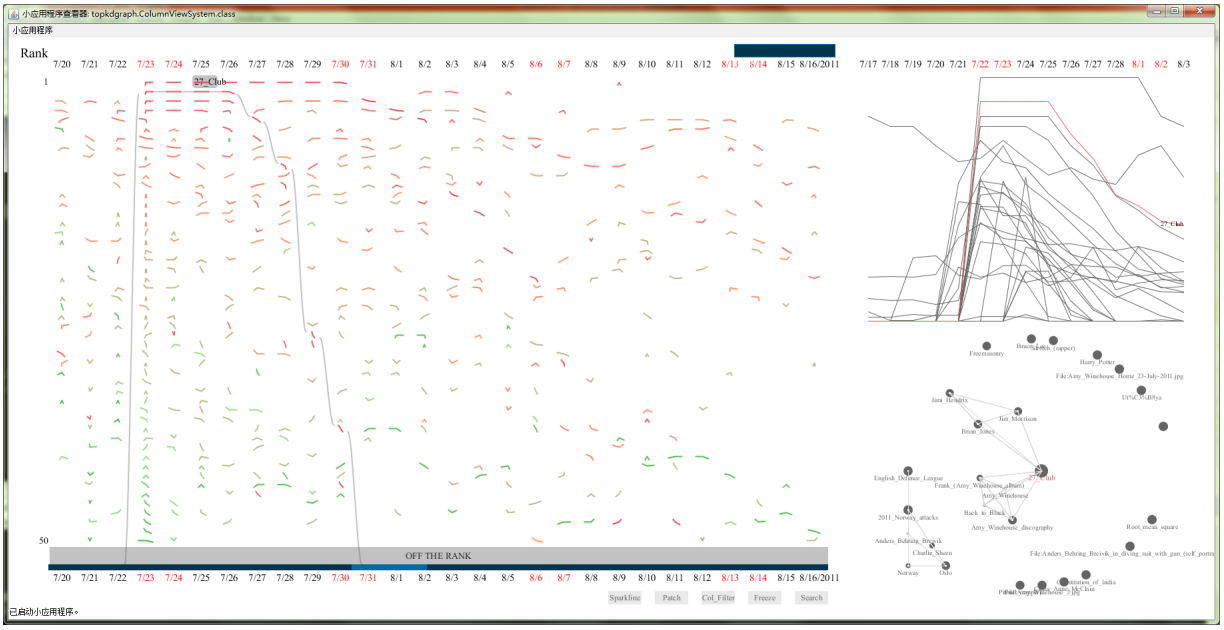

Figure 7: .

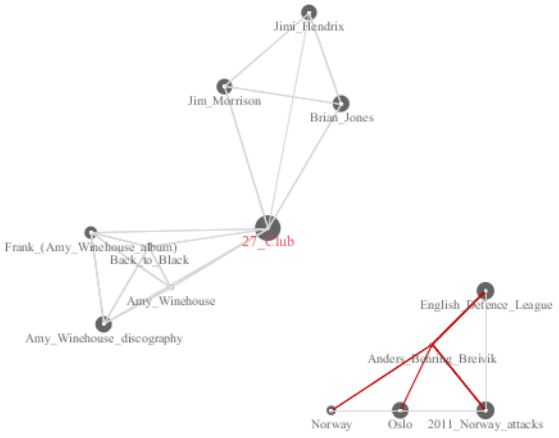

Figure 8: .

rock/popular musicians who died at age 27. And the items in N12 are names of whom belonging to 27 Club. On July 23rd 2011, Amy, the famous British rock musician, dead at her age of 27. (Event 1) So that?s why there was a network centered with 27 Club, consisting of a sub-network about Amy (N11) and N12 about other ?27 Club members?.

As for N2, the online news engine told us that on July 23rd 2011, a shooting accident conducted by Anders happened in Oslo, Norway, which is know as 2011 Norway Attack. (Event 2)

Thus, WikiTopReader could wisely spot events and efficiently help to recognize concurrent events based on multi-viewed analysis.

5.2 TV Drama Pattern?

6 User Evaluation

6.1 Experimental Conditions

6.2 Discussions

7 Conclusions

Visualizing ranking trends of Wikipedia top viewed pages is a challenging task, let alone understanding the correlation between pages. In this paper we proposed three glyph representations of Wikipedia page ranking trends, which not only avoid visual clutter but also maintain the continuity of involving ranking trends. In terms of understanding, we constructed semantic networks for pages with similar ranking trends and characterized the connections with page link information. Based on the WikiTopReader system, we demonstrated the three visual designs and the visual exploration with cases and a user study. Both the cases and the user study verify its efficiency in detecting involving ranking patterns and page-wise correlations of pages.

For future work, we plan to practice the application on real streaming data updating daily reports of Wikipedia top query. We also want to enhance the semantic network by explaining how the two pages are linked together.

References

[1] Conglei Shi, Weiwei Cui, Shixia Liu, Panpan Xu, Wei Chen, and Huamin Qu. Rankexplorer: Visualization of ranking changes in large time series data. *IEEE Transactions on Visualization and Computer Graphics*, pages 2669–2678, 2012.

[2] David Milne and Ian H Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[3] Edward R Tufte, and PR Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[4] Evgeniy Gabrilovich, and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

[5] Jishang Wei, Zeqian Shen, Nee Sundaresanl, and Kwan-Liu Ma. Visual cluster exploration of web clickstream data. In *Visual Analytics*

Science and Technology (VAST), 2012 IEEE Conference on, pages 3–12. IEEE, 2012.

[6] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

[7] Michael Batty. Rank clocks. *Nature*, 444:592–597, 2006.

[8] Mingxuan Sun, Guy Lebanon, and Kevyn Collins-Thompson. Visualizing differences in web search algorithms using the expected weighted hoeffding distance. In *Proceedings of the 19th international conference on World Wide Web*, pages 931–940, New York, NY, USA, 2010.

[9] Paul Kidwell, Guy Lebanon, and William Cleveland. Visualizing incomplete and partially ranked data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1356–1363, 2008.

[10] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. Lineup: Visual analysis of multi-attribute rankings. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2277–2286, 2013.

[11] Xi C Chen, Abdullah Mueen, Vijay K Narayanan, Nikos Karampatziakis, Gagan Bansal, and Vipin Kumar. Online discovery of group level events in time series. *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 632–640, 2014.